UIJRT ISSN: 2582-6832
United International Journal for Research & Technology
Volume 06, Issue 02, 2024 | Open Access | ISSN: 2582-6832

Modified Support Vector Machine Algorithm for Text Classification Applied in Psychiatric Tele-Triage

Samantha Vivien L. Servo[1] and Kelly Denise A. Inso[2]
[1,2]Student, College of Information Systems and Technology Management, Pamantasan ng Lungsod ng Maynila (University of the City of Manila), Manila Philippines

***Abstract*—** This study investigates the use of Support Vector Machine (SVM) models to enhance text classification for tele-triage in psychiatry. The issue addressed is SVM's tendency to ignore significant textual features, which results in low precision and recall, particularly in multi-class classification tasks with imbalanced classes. In order to address this, the researchers propose generating embeddings using the Large Language Model (LLM) RoBERTa, then reducing the dimensionality using PCA before training the SVM model. The dataset includes 500 Reddit posts with five categories of suicide risk: Attempt, Behavior, Ideation, Indicator and Supportive. Experts used the Columbia Suicide Severity Rating Scale (C-SSRS) to sort these posts. Results show significant improvement over the baseline SVM model. The model initially had trouble with recall and precision, especially for the Attempt class, which had zero precision. Significant gains were observed in the Supportive class (precision: 0.55 to 0.59, recall: 0.43 to 0.57) and Behavior (precision: 0.25 to 0.31, recall: 0.13 to 0.27) following the implementation of the RoBERTa-based strategy. Even though the attempt demonstrated some improvement (precision: 0.00 to 0.33), more optimization is required. These results suggest that incorporating RoBERTa embeddings and PCA for dimensionality reduction can enhance SVM's performance by preventing the loss of important features. The model still has issues with minority classes, suggesting that more research is needed to enhance recall for underrepresented categories and handle class imbalances.

***Keywords*—** Large Language Model, Multiclass classification, Support Vector Machine, Tele-triage, Text Classification.

I. INTRODUCTION

Support Vector Machine (SVM), an algorithm widely known for its ability to create a hyperplane decision boundary to classify data in a high-dimensional feature space, relies on support vectors to determine this boundary effectively (Kavi & Pon, 2023). SVM comes in two main types: Linear and Nonlinear. The Nonlinear classification method employs the "kernel trick" to map inputs into higher-dimensional spaces, facilitating the creation of optimal margins. These margins are designed to maximize the separation between classes, ultimately minimizing classification errors (Mahesh, 2020).

Support Vector Machine (SVM) operates by choosing specific subsets of training samples that represent distinctive characteristics. This process ensures that classifying these subsets is equivalent to dividing the entire dataset. With this, SVM is widely applied and utilized to solve different classification problems such intrusion detection, classification of facial expression, prediction of time series, speech recognition, image recognition, signal processing, detection of genes, text classification, recognition of fonts, diagnosis of faults, chemical analysis, recognition of images and other fields (Abdullah & Abdulazeez, 2021). According to Pisner and Schnyer (2020), SVM is also extensively utilized for classification in the medical field due to its high versatility across a variety of data science scenarios possible. With this in mind, according to an article entitled, "Telehealth, Telemedicine, and Telecare: What's What?" (2022), telehealth includes a broader scope in terms of remote healthcare services beyond the doctor-patient relationship. On the other hand, telemedicine is defined as the usage of telecommunications technologies to support delivery of all kinds of medical, diagnostic, and treatment-related services by doctors. Therefore, the research aims to improve the existing telemedicine application by implementing a chatbot which aims to be a mental health virtual tele-triaging. In general terms, a triage is a protocol done to screen patients seeking virtual services to prevent underestimation of severity of illness, sort patients to place of service, and determine if a need exists to escalate to an in-person evaluation or higher level of care (Kobeissi & Ruppert, 2021).

However, despite the extensive applications of SVM, it still encounters various challenges. Richariya et al. (2020) noted that SVM struggles with understanding data distribution, which greatly affects its ability to generalize effectively and avoid overfitting. To address this, Large Language Model (LLM) was introduced to

UIJRT ISSN: 2582-6832
United International Journal for Research & Technology
Volume 06, Issue 02, 2024 | Open Access | ISSN: 2582-6832

the algorithm to prioritize essential data for classification. This model was chosen by the researchers as it is a machine that uses deep learning to understand, generate, and predict human language that is primarily needed for this study.

A. Statement of the Problem

Support Vector Machine (SVM) algorithm faces a limitation in text classification tasks due to their tendency to discard essential features of textual data.

SVM does not favor distribution of data due to lack of knowledge, which can result in the elimination of important features (Richariya et. al., 2020).

Moreover, flaws in textual data, such as incompleteness or noise, can significantly affect the success of text classification using SVM (Behzadidoost et. al., 2024). This is because the algorithm's performance depends heavily on the quality and complexity of the data. (Ding et. al., 2021).

	precision	recall	f1-score	support
Behavior	0.25	0.13	0.17	15
Supportive	0.55	0.43	0.48	28
Indicator	0.35	0.32	0.33	19
Attempt	0.00	0.00	0.00	13
Ideation	0.26	0.56	0.36	25
accuracy			0.34	100
macro avg	0.28	0.29	0.27	100
weighted avg	0.32	0.34	0.31	100

Figure 1: *Classification Report after applying the baseline SVM to the Reddit C-SSRS Suicide Dataset*

B. Objective of the Study

The objective of the study is to effectively represent textual data and address its incompleteness. By providing adequate representation of textual data such as parameters and penalty factors, there's an improvement of text classification performance in Support Vector Machine algorithm (Tao et. al., 2019). Highlighted in the study of Gul et. al. (2020), the validity of outcome data will also be a factor when representation is present. The larger the value of the data, the greater the authenticity of the data is affected since it highly depends on the representation placed in the algorithm. Additionally, it is stated by Gasparetto et. al. (2022) that representation in SVM is fundamental to achieving good performance on classification.

II. METHODOLOGY

A. Requirement Analysis

In modifying Support Vector Machine Algorithm for Text Classification applied in a Tele-Triage, two key requirements are addressed. The first requirement focuses on ensuring the algorithm classifying imbalance of data by implementing generated RoBERTa embeddings in the algorithm. This aims to not lead to biased predictions and poor generalization that will lead to negative performance of the algorithm since RoBERTa embeddings capture context-dependent meanings. This feature allows the model to understand the language more like how a human being does. The second requirement focuses on reducing dimensionality before SVM model training with the help of Principal Component Analysis (PCA). This aims to make analysis more manageable while retaining meaningful variance that is needed for text classification. Since with the PCA it discards lower-variance components that primarily removes noise within the data. In addition, it simplifies the data structure that mitigates the curse of dimensionality. With these two requirements successfully applied to the traditional algorithm, this will greatly enhance the algorithm's effectiveness of text classification not only in a tele-triage setting but also in general usage.

B. Research Design

Use of Large Language Model (RoBERTa)

The use of Large Language Model (LLM) specifically Robustly Optimized BERT Pretraining Approach (RoBERTa) in the study will be beneficial on making the algorithm effective once applied to a Tele-Triage. Since, RoBERTa is a reimplementation of BERT that allows grasp the language complexities by modifying pretraining process leading on natural language processing (NLP) tasks. The usage of RoBERTa instead of BERT primarily centers on the feature of training

Volume 06, Issue 02, 2024 | Open Access | ISSN: 2582-6832

more data that will improve performance and better generalization.

C. Methods and Performance Metric

The flowchart derives from Okur, et al. (2024)'s proposed methodology in which the flowchart consists of four (4) nodes. The first node consists of text preprocessing in which data will be input to first undergo text preprocessing. The data used in the study is from Gaur et al. 2019 "Reddit C-SSRS Suicide Dataset". The text preprocessing techniques used for the dataset includes tokenization and stemming. The second node is where the preprocessed data will undergo text vector representation by implementing RoBERTa embeddings which is used to solve the objective of the study. The third node is classification using the enhanced SVM and lastly, assessing the classification results using the metrics.

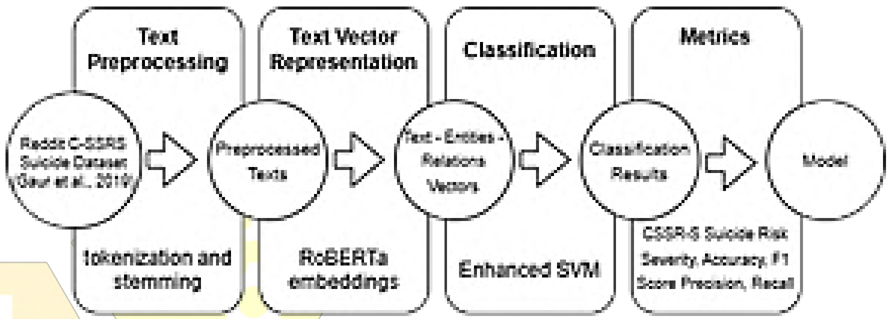

Figure 2: *Conceptual Framework for the modification of Support Vector Machine in Text Classification*

Suicide Risk Severity Lexicon

Suicide Rish Severity Lexicon consists of five (5) Classification:

1. Attempt
2. Behavior
3. Ideation
4. Indicator
5. Supportive

This lexicon is derived from the Columbia Suicide Rating Scale (C-SSRS) that is primarily used in the study as a basis for the researchers to assess the underlying mental health issues the users are facing.

F1-Score

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-Score is a metric under confusion matrix that is used to evaluate the performance of a classification model. This metric was primarily used throughout the study as it is made for focusing on the imbalance of data that the study mainly aims to solve. The researchers will assess the results of the metrics by testing whether the results range from 0 and 1. Wherein, 1 is equivalent to a perfect performance where no false positives and no false negatives are present. Meanwhile, 0 is equivalent to a worst performance where the model fails to provide a correct prediction of the data.

III. RESULTS AND DISCUSSION

This chapter presents the findings from the proposed enhanced SVM algorithm. The baseline SVM model showed poor performance in several important areas. The precision for the Attempt class is 0.00 and recall for the Behavior class is extremely low at 0.13. These figures resulted in overall low F1-scores for several classes, including a macro average F1-score of 0.27 and a weighted average F1-score of 0.31. In contrast, the enhanced SVM model showed better performance. The overall accuracy increased from 0.34 to 0.39, with improvements across precision, recall, and F1-score for several classes. The Supportive class had improved in both precision (from 0.55 to 0.57) and recall (from 0.43 to 0.61), resulting in a higher F1-score (0.59). In addition, the Behavior class had improved its precision from 0.25 to 0.42, and recall from 0.13 to 0.33. Overall, the F1-score had a notable improvement from 0.17 to 0.37. However, despite these improvements, the Attempt class continues to present challenges, with its precision improving slightly from 0.00 to 0.33, but recall remaining low at 0.08.

Classification Report:

	precision	recall	f1-score	support
Attempt	0.33	0.08	0.12	13
Behavior	0.42	0.33	0.37	15
Ideation	0.26	0.44	0.32	25
Indicator	0.42	0.26	0.32	19
Supportive	0.57	0.61	0.59	28
accuracy			0.39	100
macro avg	0.40	0.34	0.35	100
weighted avg	0.41	0.39	0.38	100

Figure 3: *Classification Report after applying the improved SVM to the Reddit C-SSRS Suicide Dataset*

IV. CONCLUSION

The results suggest that by incorporating RoBERTa embeddings and PCA for dimensionality reduction, it can enhance the SVM's performance by preventing the loss of important features. However, while feature selection and class balancing techniques helped improve the model, further work is needed to address some of the challenges from classifying the underrepresented class.

V. RECOMMENDATION

Based on the results, the model works well with the Supportive and Behavior classes, but still has issues with minority classes. This suggests more research regarding the enhancement of recall for the underrepresented categories (such as the Attempt class) is needed to handle class imbalances. Application of other types of feature selection techniques can also help improve the model further. These recommendations are essential for advancing the application of Support Vector Machine algorithm on multiclass text classification problems.

REFERENCES

[1] Behzadidoost, R., Mahan, F., & Izadkhah, H. (2024). Granular computing-based deep learning for text classification. Information Sciences, 652, 119746. https://doi.org/10.1016/j.ins.2023.119746

[2] Ding, X., Yang, F., Jin, S., & Cao, J. (2021). An efficient alpha seeding method for optimized extreme learning machine-based feature selection algorithm. Computers in Biology and Medicine, 134, 104505. https://doi.org/10.1016/j.compbiomed.2021.104505

[3] Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. Information, 13(2), 83. https://doi.org/10.3390/info13020083

[4] Gul, E., Alpaslan, N., & Emiroglu, M. E. (2021). Robust optimization of SVM hyper-parameters for spillway type selection. Ain Shams Engineering Journal, 12(3), 2413–2423. https://doi.org/10.1016/j.asej.2020.10.022

[5] Kavi, P. S., & Pon, K. K. (2023, November). An embedded feature selection approach for depression classification using short text sequences. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1568494623008463

[6] Kobeissi, M. M., & Ruppert, S. D. (2021). Remote patient triage: Shifting toward safer telehealth practice. Journal of the American Association of Nurse Practitioners, 34(3), 444–451. https://doi.org/10.1097/jxx.0000000000000655

[7] Mahesh, Batta. (2019). Machine Learning Algorithms -A Review. ResearchGate. 10.21275/ART20203995.

[8] Mustafa Abdullah, D., & Mohsin Abdulazeez, A.. (2021). Machine Learning Applications based on SVM Classification A Review. Qubahan Academic Journal, 1(2), 81–90. https://doi.org/10.48161/qaj.v1n2a50

[9] Okur, H. I., Tohma, K., & Sertbas, A. (2024). Relational Turkish text classification using distant supervised entities and relations. Computers, Materials & Continua, 79(2), 2209–2228. https://doi.org/10.32604/cmc.2024.050585

[10] Pisner, D. A., & Schnyer, D. M. (2020). Support Vector Machine. Machine Learning, 101–121. https://doi.org/10.1016/b978-0-12-815739-8.00006-7

[11] Richhariya, B., & Tanveer, M. (2020). Universum least squares twin parametric-margin support vector machine. 2020 International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/ijcnn48605.2020.9206865

[12] Tao, Z., Huiling, L., Wenwen, W., & Xia, Y. (2019). Ga-SVM based feature selection and parameter optimization in hospitalization expense modeling. Applied Soft Computing, 75, 323–332. https://doi.org/10.1016/j.asoc.2018.11.001

[13] Telehealth, Telemedicine, and Telecare: What's What? (2022). Fcc.gov. https://www.fcc.gov/general/telehealth-telemedicine-and-telecare-whats-what